Phân tích gom nhóm

Gom nhóm

- Có rất nhiều ứng dụng yêu cầu việc phân hoạch các điểm dữ liệu thành các nhóm gồm những điểm giống nhau.

- Việc phân hoạch một lượng lớn dữ liệu thành một số nhỏ hơn rất có ích cho nhiều ứng dụng khai phá dữ liệu.

Gom nhóm

- Một số ứng dụng tiêu biểu của việc gom nhóm.
  - Tóm tắt dữ liệu.
  - Chia nhóm khách hàng.
  - Phân tích mạng xã hội.
  - Hỗ trợ các ứng dụng khai phá dữ liệu khác.

Chọn đặc trưng để gom nhóm

- Mục tiêu chính của việc chọn đặc trưng là để loại bỏ các thuộc tính nhiễu.
- Việc chọn lọc đặc trưng thường khó hơn với các bài toán không giám sát như gom cụm do thiếu các tiêu chí đánh giá ngoài như nhãn dữ liệu.
- Có 2 lớp mô hình chính để chọn đặc trưng
  - Mô hình lọc
  - Mô hình gói

Mô hình lọc

- Với các mô hình lọc, một tiêu chí cụ thể được dùng để đánh giá ảnh hưởng của các đặc trưng hoặc nhóm đặc trưng cụ thể.
- Chúng ta có một số tiêu chí phổ biến như sau.
  - Term strength.
  - Sự phụ thuộc thuộc tính dự đoán.
  - Entropy
  - Thống kê Hopkins

Mô hình lọc

- Term strength is suitable for sparse domains such as text data. In such domains, it is more meaningful to talk about presence or absence of nonzero values on the attributes (words), rather than distances. Furthermore, it is more meaningful to use similarity functions rather than distance functions. In this approach, pairs of documents are sampled, but a random ordering is imposed between the pair. The term strength is defined as the fraction of similar document pairs (with similarity *greater* than β), in which the term occurs in both the documents, conditional on the fact that it appears in the first. In other words, for any term t, and document pair $(\overline{X}, \overline{Y})$ that have been deemed to be sufficiently similar, the term strength is defined as follows:

$$\text{Term Strength} = P(t \in \overline{Y} | t \in \overline{X}). \tag{6.1}$$

If desired, term strength can also be generalized to multidimensional data by discretizing the quantitative attributes into binary values. Other analogous measures use the correlations between the overall distances and attribute-wise distances to model relevance.

Source: Data Mining, Charu C. Aggarwal

Mô hình lọc

- Sự phụ thuộc thuộc tính dự đoán.

The intuitive motivation of this measure is that correlated features will always result in better clusters than uncorrelated features. When an attribute is relevant, other attributes can be used to predict the value of this attribute. A classification (or regression modeling) algorithm can be used to evaluate this predictiveness. If the attribute is numeric, then a regression modeling algorithm is used. Otherwise, a classification algorithm is used. The overall approach for quantifying the relevance of an attribute i is as follows:

1. Use a classification algorithm on all attributes, except attribute i, to predict the value of attribute i, while treating it as an artificial class variable.

2. Report the classification accuracy as the relevance of attribute i.

Any reasonable classification algorithm can be used, although a nearest neighbor classifier is desirable because of its natural connections with similarity computation and clustering. Classification algorithms are discussed in Chap. 10.

Source: Data Mining, Charu C. Aggarwal

Mô hình lọc

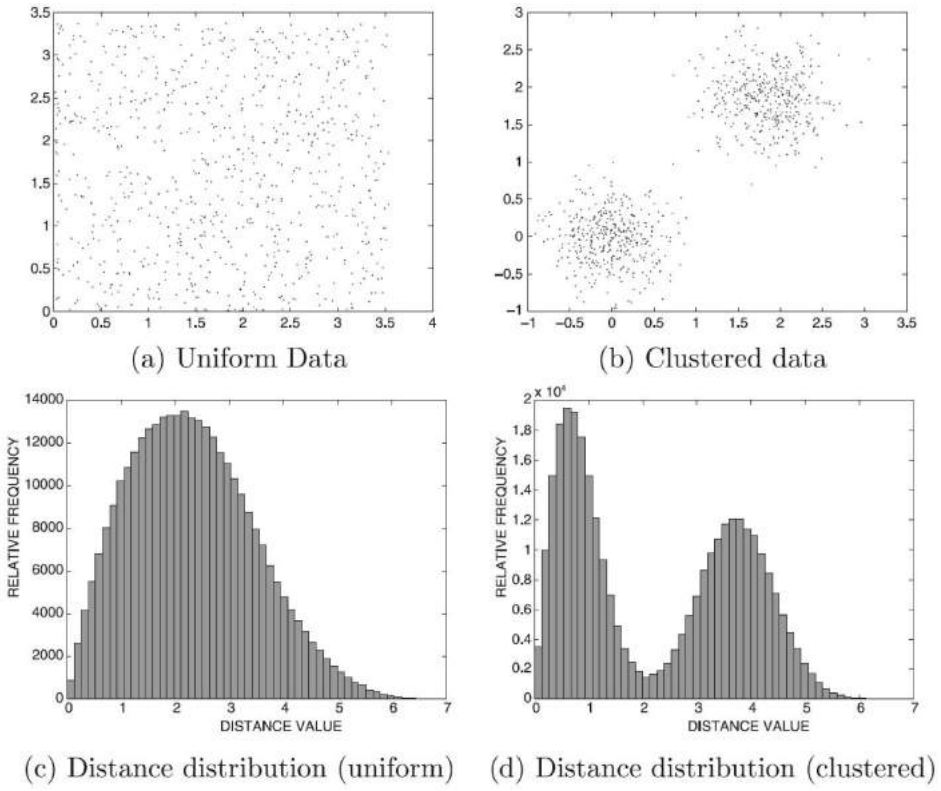
Entropy

- Ý tưởng chính ở đây là việc các dữ liệu có nhóm thể hiện các đặc trưng gom nhóm trên phân phối khoảng cách tương ứng.
- Mục tiêu của các độ đo entropy là định lượng hóa hình dạng của phân phối khoảng cách với một tập con các đặc trưng.

Chọn đặc trưng để gom nhóm

Mô hình lọc

Entropy



(a) Uniform Data (b) Clustered data

(c) Distance distribution (uniform) (d) Distance distribution (clustered)

Figure 6.1: Impact of clustered data on distance distribution entropy

Source: Data Mining, Charu C. Aggarwal

Mô hình lọc

Thống kê Hopkins.

Source: Data Mining, Charu C. Aggarwal

Let $\mathcal{D}$ be the data set whose clustering tendency needs to be evaluated. A sample S of r synthetic data points is randomly generated in the domain of the data space. At the same time, a sample R of r data points is selected from $\mathcal{D}$. Let $\alpha_1 \ldots \alpha_r$ be the distances of the data points in the sample $R \subseteq D$ to their nearest neighbors within the original database $\mathcal{D}$. Similarly, let $\beta_1 \ldots \beta_r$ be the distances of the data points in the synthetic sample S to their nearest neighbors within $\mathcal{D}$. Then, the Hopkins statistic H is defined as follows:

$$H = \frac{\sum_{i=1}^{r} \beta_i}{\sum_{i=1}^{r} (\alpha_i + \beta_i)}. \tag{6.3}$$

The Hopkins statistic will be in the range $(0, 1)$. Uniformly distributed data will have a Hopkins statistic of 0.5 because the values of α_i and β_i will be similar. On the other hand, the values of α_i will typically be much lower than β_i for the clustered data. This will result in a value of the Hopkins statistic that is closer to 1. Therefore, a high value of the Hopkins statistic H is indicative of highly clustered data points.

Mô hình gói

- Các mô hình gói sử dụng một tiêu chí đánh giá gom nhóm bên trong cùng với một thuật toán gom nhóm áp dụng lệ một tập các đặc trưng thích hợp.

- Ý tưởng ở đây là dùng một thuật toán gom nhóm với một tập các đặc trưng và sau đó đánh giá chất lượng với tiêu chí được chọn.

- Một khuyết điểm lớn của cách này là sự nhạy với tiêu chí đánh giá được chọn.

Mô hình gói

- Một cách khác đơn giản hơn là chọn các đặc trưng riêng lẻ với một tiêu chí chọn đặc trưng của các thuật toán phân loại.

1. Use a clustering algorithm on the current subset of selected features F, in order to fix cluster labels L for the data points.
2. Use any *supervised* criterion to quantify the quality of the individual features with respect to labels L. Select the top-k features on the basis of this quantification.

Source: Data Mining, Charu C. Aggarwal

Mô hình gói

Algorithm *GenericRepresentative*(Database: $\mathcal{D}$, Number of Representatives: k)
begin
  Initialize representative set S;
  repeat
    Create clusters $(\mathcal{C}_1 \ldots \mathcal{C}_k)$ by assigning each
      point in $\mathcal{D}$ to closest representative in S
      using the distance function $Dist(\cdot, \cdot)$;
    Recreate set S by determining one representative $\overline{Y_j}$ for
      each $\mathcal{C}_j$ that minimizes $\sum_{\overline{X_i} \in \mathcal{C}_j} Dist(\overline{X_i}, \overline{Y_j})$;
  until convergence;
  return $(\mathcal{C}_1 \ldots \mathcal{C}_k)$;
end

Figure 6.2: Generic representative algorithm with unspecified distance function

Source: Data Mining, Charu C. Aggarwal

Các thuật toán dựa theo đại diện

- Các thuật toán dựa theo đại diện dựa trực tiếp vào khái niệm khoảng cách (hoặc sự tương đồng) để gom nhóm các điểm dữ liệu.

- Trong các thuật toán này, các nhóm được tạo trong một lần và không có quan hệ phân tầng với các nhóm.

Typically, it is assumed that the number of clusters, denoted by k, is specified by the user. Consider a data set $\mathcal{D}$ containing n data points denoted by $\overline{X_1} \ldots \overline{X_n}$ in d-dimensional space. The goal is to determine k representatives $\overline{Y_1} \ldots \overline{Y_k}$ that minimize the following objective function O:

$$O = \sum_{i=1}^{n} \left[\min_j Dist(\overline{X_i}, \overline{Y_j})\right]. \quad (6.4)$$

Source: Data Mining, Charu C. Aggarwal

Các thuật toán dựa theo đại diện

- Các bài toán tối ưu này thường được giải bằng phương pháp lặp với mỗi bước lặp có 2 bước con.

- (Assign step) Assign each data point to its closest representative in S using distance function $Dist(\cdot, \cdot)$, and denote the corresponding clusters by $\mathcal{C}_1 \ldots \mathcal{C}_k$.
- (Optimize step) Determine the optimal representative $\overline{Y_j}$ for each cluster $\mathcal{C}_j$ that minimizes its *local* objective function $\sum_{\overline{X_i} \in \mathcal{C}_j} [Dist(\overline{X_i}, \overline{Y_j})]$.

Source: Data Mining, Charu C. Aggarwal

Các thuật toán dựa theo đại diện

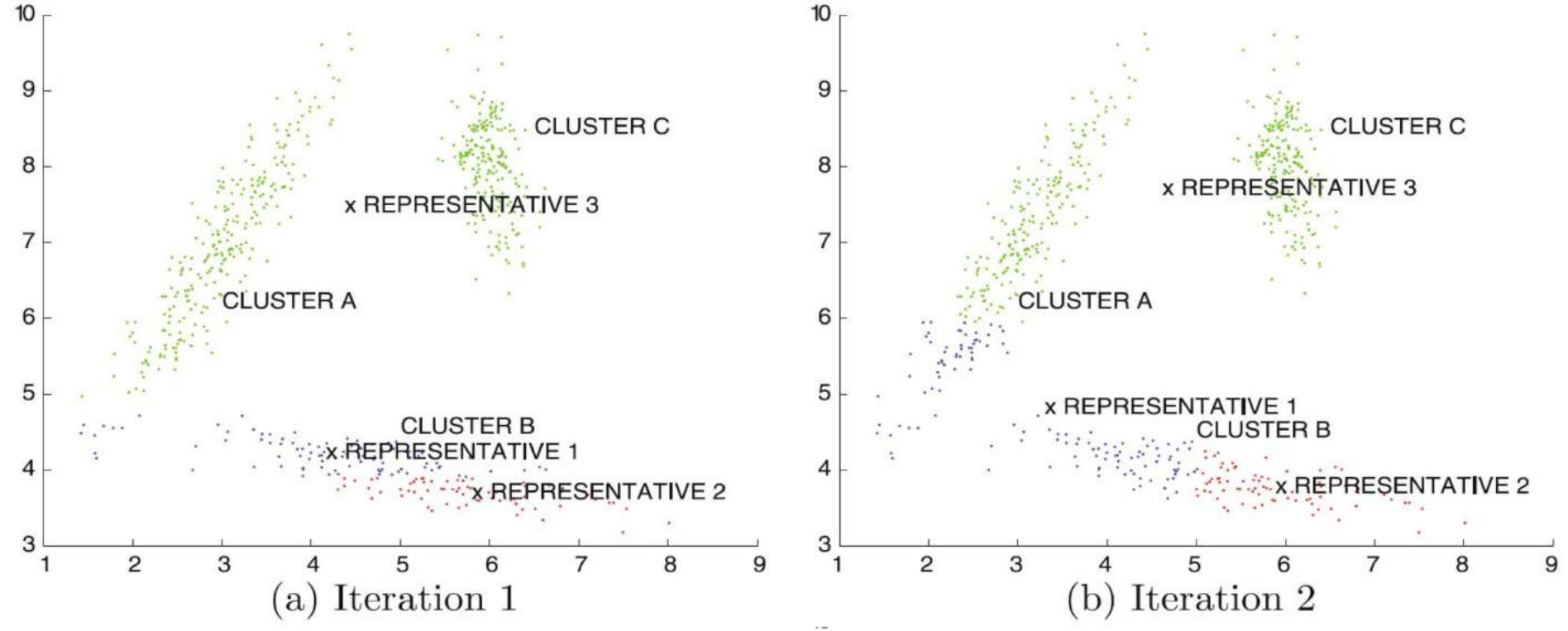
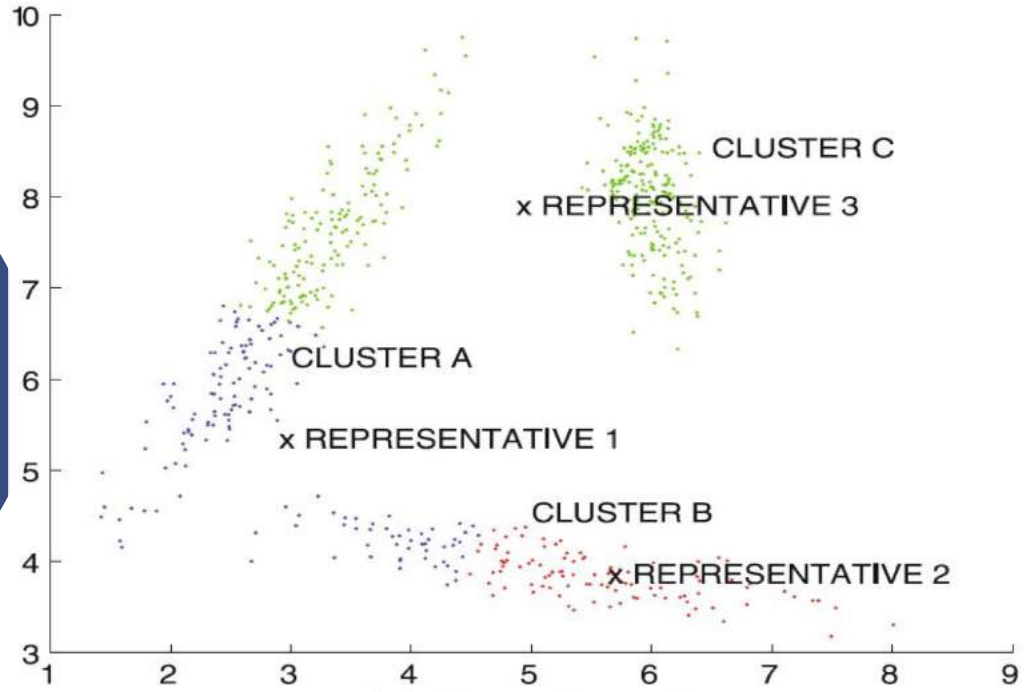


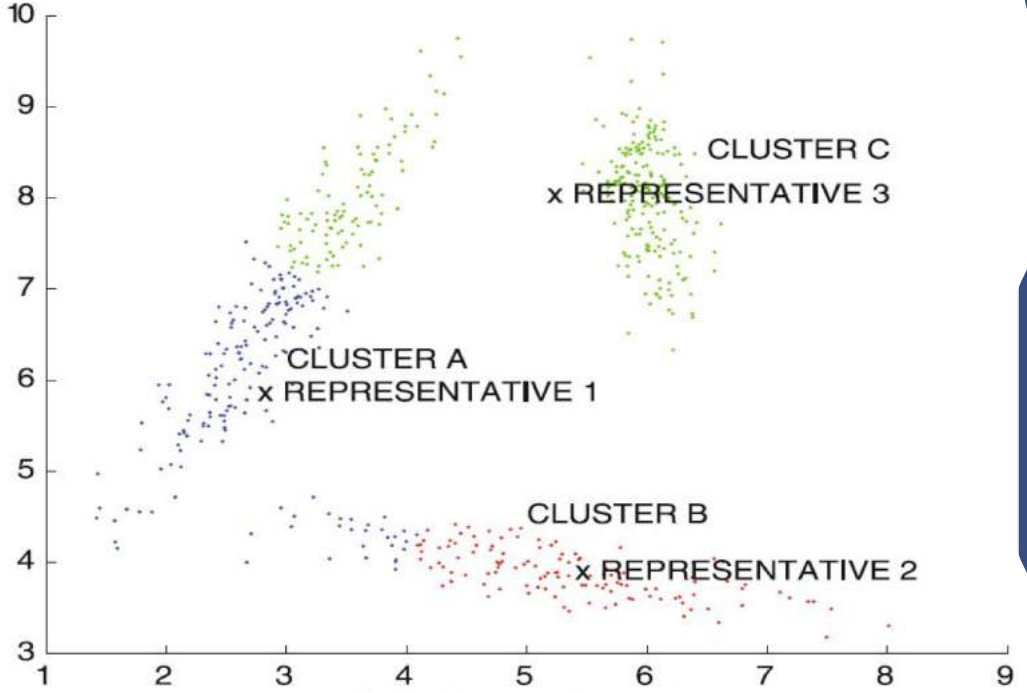
Figure 6.3: Illustration of k-representative algorithm with random initialization

Source: Data Mining, Charu C. Aggarwal

Các thuật toán dựa theo đại diện



(c) Iteration 3

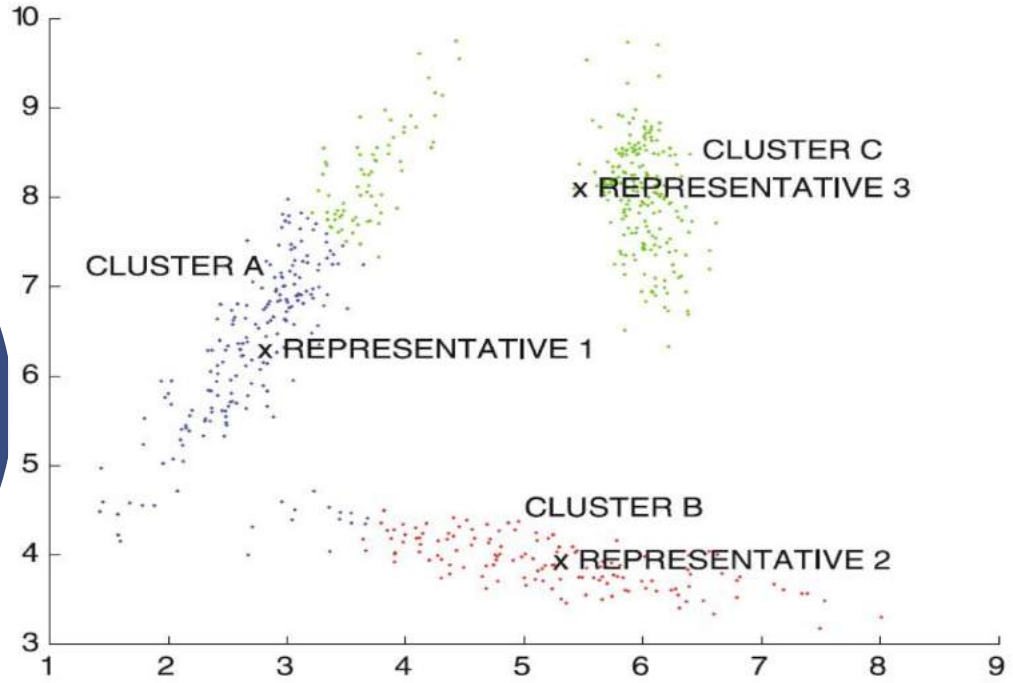


(d) Iteration 4

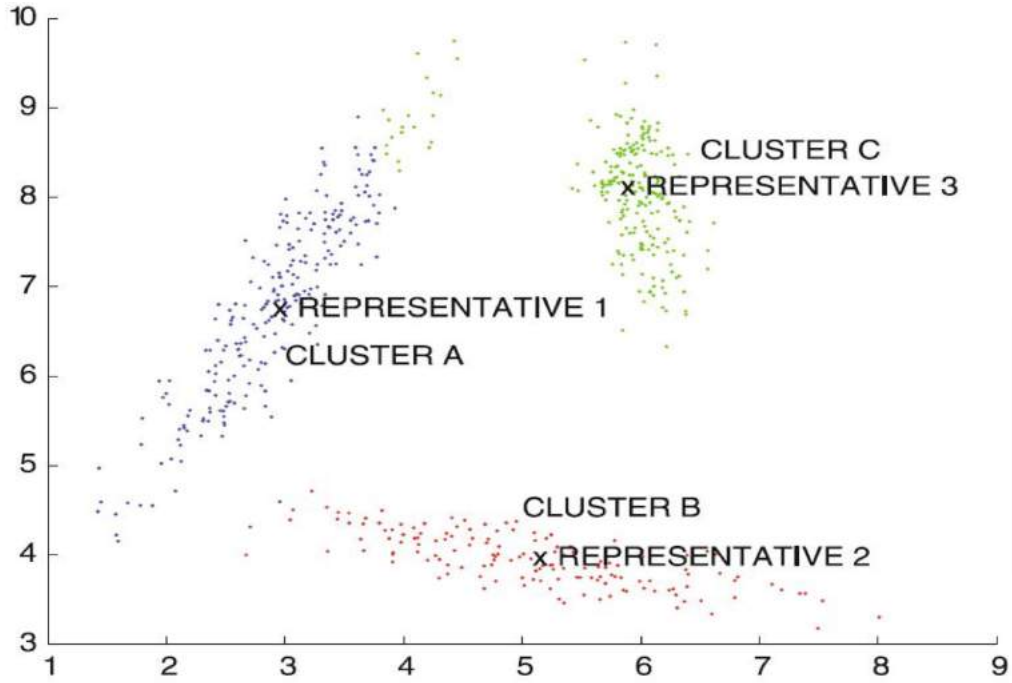
Figure 6.3: Illustration of k-representative algorithm with random initialization

Source: Data Mining, Charu C. Aggarwal

Các thuật toán dựa theo đại diện



(e) Iteration 5



(f) Iteration 10

Figure 6.3: Illustration of k-representative algorithm with random initialization

Source: Data Mining, Charu C. Aggarwal

Các thuật toán dựa theo đại diện

Thuật toán k-Means

- Trong các thuật toán k-means, tổng bình phương khoảng cách từ các điểm dữ liệu đến đại diện gần nhất được dùng cho hàm mục tiêu của bài toán tối ưu.

$$Dist(\overline{X_i}, \overline{Y_j}) = ||\overline{X_i} - \overline{Y_j}||_2^2.$$

Source: Data Mining, Charu C. Aggarwal

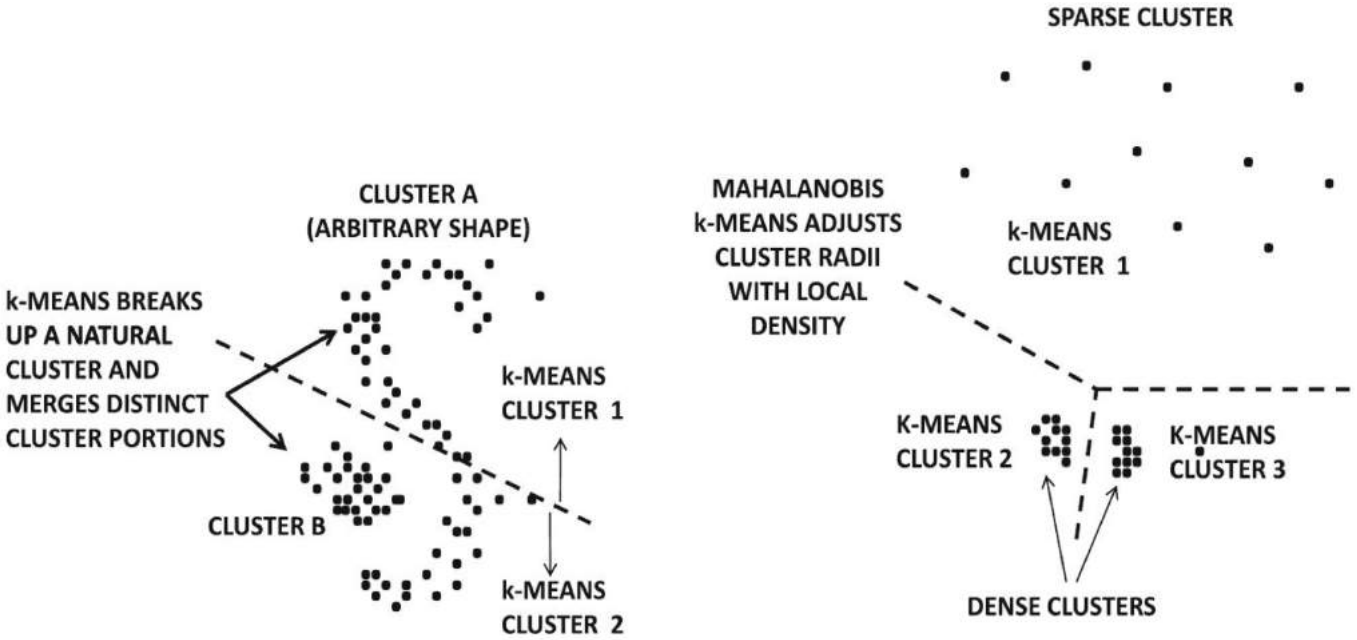
Thuật toán k-Means

- Ngoài ra, chúng ta còn có biến thể khác sử dụng khoảng cách Mahalanobis

$$Dist(\overline{X_i}, \overline{Y_j}) = (\overline{X_i} - \overline{Y_j})\Sigma_j^{-1}(\overline{X_i} - \overline{Y_j})^T.$$

Các thuật toán dựa theo đại diện

Thuật toán k-Means



(a) Varying cluster shape (Bad for k-means)

(b) Varying cluster density (Good for Mahalanobis k-means)

Figure 6.4: Strengths and weaknesses of k-means

Source: Data Mining, Charu C. Aggarwal

Thuật toán kernel k-Means

- Thuật toán k-mean có thể được mở rộng để tìm các cụm với hình dạng bất kì bằng kĩ thuật kernel

- Ý tưởng cơ bản là biến đổi ẩn dữ liệu sao cho các cụm với hình dạng bất kì được nối đến các cụm Euclid trong không gian mới.

Thuật toán k-Medoids

- Thuật toán k-medoids mặc dù cũng dùng khái niệm đại diện và hàm mục tiêu, nhưng có cấu trúc thuật toán khác các phương pháp đã đề cập.

- Khác biệt chính của thuật toán này là các đại diện luôn được chọn từ các phần tử trong cơ sở dữ liệu.

Thuật toán k-Medoids

Algorithm *GenericMedoids*(Database: $\mathcal{D}$, Number of Representatives: k)
begin
  Initialize representative set S by selecting from $\mathcal{D}$;
  repeat
    Create clusters $(\mathcal{C}_1 \ldots \mathcal{C}_k)$ by assigning
      each point in $\mathcal{D}$ to closest representative in S
      using the distance function $Dist(\cdot, \cdot)$;
    Determine a pair $\overline{X_i} \in \mathcal{D}$ and $\overline{Y_j} \in S$ such that
     replacing $\overline{Y_j} \in S$ with $\overline{X_i}$ leads to the
     greatest possible improvement in objective function;
    Perform the exchange between $\overline{X_i}$ and $\overline{Y_j}$ only
     if improvement is positive;
  until no improvement in current iteration;
  return $(\mathcal{C}_1 \ldots \mathcal{C}_k)$;
end

Figure 6.5: Generic k-medoids algorithm with unspecified hill-climbing strategy

Source: Data Mining, Charu C. Aggarwal